Instructions

Please upload your project report as a file attachment. The title page should include the title of the project, name of the participants, and a brief abstract. The rest of the report should be typed and double-spaced on standard-sized paper, with 1" margins on all sides. Your report should be approximately 15 pages (with a margin of  pages). If you think it is not possible to fit the material within this page limit, please contact me.

The report should have an introduction with clear description of the problem and (if applicable) references to relevant work on that topic. There should be a methodology section with concise description of the techniques used in the data analysis. This should be followed by the result section where you should discuss the most significant findings (with appropriate tables and plots) of your analysis. If you plan to use multiple algorithms, please create subsections for each of them in the methodology and results sections. The report should conclude with a conclusion section where main takeaways, limitation of the analysis and future direction should be discussed briefly.

Please do not include any R output in the main part of the report. If you want to show results obtained from a R output, please create tables for them to include in the main report. This instruction, however, is not valid for plots. You may put R plots in the main body of the report. Please create a supplementary for the R outputs. The supplementary does not have any page limit. If you want me to look for specific results in the supplementary section, please reference that appropriately in the main report.

A Cluster-Based Approach to Fitting Regression Models

L. Austin Hadamuscin, James Hawkins, Tomiwa Omotesho, Deep Sagar Karki

STAT 6440: Data Mining

Dr. Shuchismita Sarkar

April 23, 2022

Abstract: Many have anxiety about dying early, others fear a long life of pain, but what factors lead to a long healthy life, and are those factors the same around the world? This paper attempts model healthy life expectancy using data retrieved from the World Health Organization (WHO). k-NN is used as a method of data imputation for missing data. Model Based and k-Means clustering were considered as clustering methods to split the countries into an unknown number of clusters. The ensemble method, random forest, was used to determine the important factors in each cluster and each cluster LASSO regression to model healthy life expectance at birth. Due to the relatively small size of the data set, leave-one-out cross validation (LOOCV) was used for any unknown variables that need data mined. Unsurprisingly, we found that there were two common important variables, the adult mortality rate and infant deaths; with other important variables varying across the clusters.

Contents

[1 Introduction 4](#_Toc101088286)

[1.1 Motivation 4](#_Toc101088287)

[1.2 Objectives 4](#_Toc101088288)

[2 Data Description 4](#_Toc101088289)

[3 Methodology 4](#_Toc101088290)

[3.1 Leave-One-Out Cross-Validation 4](#_Toc101088291)

[3.2 k-NN 5](#_Toc101088292)

[3.3 k-Means Clustering 6](#_Toc101088293)

[3.4 Model-Based Clustering 7](#_Toc101088294)

[3.5 Random Forest 8](#_Toc101088295)

[3.6 Linear Regression 9](#_Toc101088296)

[4 Results 11](#_Toc101088297)

[4.1 k-NN 11](#_Toc101088298)

[4.2 k-Means 11](#_Toc101088299)

[4.3 Model-Based 13](#_Toc101088300)

[4.4 K-Means vs Model-Based 15](#_Toc101088301)

[4.5 Random Forest 16](#_Toc101088302)

[4.5.1 Cluster 1 16](#_Toc101088303)

[4.5.2 Cluster 2 17](#_Toc101088304)

[6 Appendix 22](#_Toc101088305)

[7 Citations 25](#_Toc101088306)

# Introduction

## Motivation

## Objectives

# Data Description

Our variables include are:

1. Life Expectancy – life expectancy in age
   1. Healthy life expectancy (HALE) at birth (years)
   2. Healthy life expectancy (HALE) at age 60 (years)
2. Adult Mortality – mortality rate per 1000 people aged 15-60
3. Infant Deaths – Infant mortality rate (between birth and 11 months per 1000 live births)
4. Alcohol – Recorded alcohol consumption (ages 15+) per capita in liters of pure alcohol
5. Percentage Expenditure – Domestic general government health expenditure (GGHE-D) as a percentage of general government expenditure (GGE) (%)
6. Hepatitis B – Hepatitis B (HepB3) immunization coverage among 1-year-olds (%)
7. Measles – number of reported cases per 1000 population
8. BMI – Mean body mass index trends among adults, crude (kg/m²)
9. Polio – polio (Pol3) immunization coverage among 1-year-olds (%)
10. Total expenditure – Domestic general government health expenditure (GGHE-D) as a percentage of gross domestic product (GDP) (%)
11. Diphtheria – diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
12. HIV/AIDS – deaths per 1000 live births with HIV/AIDS (0-4 years)
13. Thinness 5-19 years – prevalence of thinness among adolescents age 5 to 19, BMI < -2 standard deviations below the median (crude estimate) (%)

# Methodology

## Leave-One-Out Cross-Validation

The relatively small number observations in our data allowed us the luxury of using leave-one-out cross validation (LOOCV) whenever we needed to find on optimal parameter. LOOCV uses the following steps:

* partition all data into a training set except for one observation that will be used as our testing set;
* build a model using the training set;
* use this model to predict the response value in the testing set and calculate the mean square error (MSE);
* and repeat this process until every observation has been used in the testing set.

The test MSE is then calculated by taking the average of the MSE*i* values calculated in the LOOCV process. This process is repeated using different parameters and the set of parameters with the lowest test MSE are our optimal parameters. To expedite this process, we used the caret package (Kuhn, 2021) along with other packages in the statistical programming language R (R Core Team, 2021) (RStudio Team, 2021). Using LOOCV over other methods, such as k-fold cross validation, gives us a low bias estimate of test MSE because our training set contains almost all the data; however, because substantial correlation exists between the MSE*i* from the overlapping in the training and testing sets, the variance of the estimate of test MSE is very high, meaning that any addition of new data could result in vastly different outcomes for our optimal parameters. We chose to accept this variance-bias trade-off because each observation is one country partitioning our data into a train, validation, and test set would mean that whole countries would be excluded from the model building process which greatly reduces the applicability of our final model.

## k-NN

We used k nearest neighbors (k-NN) to impute missing data. k-NN works by calculating the Euclidean distance between each observation in the data set, and each new observation, : where represents the number of predictors. In classification k-NN, the new value takes the class of the majority among its closest observations based on Euclidean distance. In the case of ties, there are several methods used to classify the new observations such as using the majority class among all observations. Our response variable is continuous, so we used regression k-NN where new observations are assigned the average response of the closest observations. We used this concept to impute our missing data. We data mined the optimal by utilizing the train function from the caret R package (Kuhn, 2021) and the knn method from the FNN R package (Beygelzimer, et al., 2019) to perform LOOCV as described in section 3.1. While doing initial research into k-NN data imputation, we found the function knnImputation from the archived R package DMwR (Torgo, 2010). We modified the code of this function to estimate the missing values for each observation, using the mean of the present values from their nearest neighbors.

## k-Means Clustering

We used a -means approach to clustering to cluster our data. The -means clustering algorithm is fairly simple and can be easily explained:

* randomly choose points in -dimentional space to use as the initial centroids;
* calculate the Euclidean distance between the observations and the centroids and assign each observation to its closest centroid;
* recompute new centroids by calculating the center of each group that was created in the previous step. Repeat the previous step;
* The algorithm stops when there is no change to the cluster assignments.

We can measure how well a -means algorithm performs by calculating its within sum of squares () which is a measure of the density of observations in the clusters and is defined as:

where is the set of observations in cluster . The performance of a -means clustering algorithm is highly dependent on the initialization of the centroids created in the first step of the algorithm so to combat this, we randomly generated a set 1000 seeds from a uniform(0,900000) distribution, using the ceiling function (R Core Team, 2021) to round to the next highest whole number. We used theses seeds to run 1000 iterations of the algorithm, choosing the seed that produced the lowest . Using the same 1000 seeds, this process was repeated for and the model that resulted in the lowest without the increase in k resulting in a negligibly lower was chosen.

## Model-Based Clustering

We also took a model-based approach to clustering our data. Model-based clustering uses finite mixture models which are a linear combination of weighted density functions:

where are the weights, , , and is the mixture order, or number of components, that use the Expected-Maximization (EM) iterative algorithm to estimate . Like the k-means clustering algorithm, model-based clustering algorithms are highly dependent on initialization so many models must be run to get the optimal assignment of clusters. To determine what set of clusters are the best, a modified version of the likelihood, Bayesian Information Criterion (BIC), is used:

where is the maximum likelihood and is the number of parameters to be estimated. We used the Mclust function from the mclust R package (Scrucca, Fop, Murphy, & Raftery, 2016) to perform the clustering. This package automatically runs the algorithm at different initializations and chooses the best model and returns its negative BIC. We ran the function using K and chose the model with the maximum negative BIC without the increase in K resulting in a negligibly higher negative BIC. We then compared this model, visually, with the model we created using k-means and chose the one that we felt created the best clusters.

## Random Forest

The necessity of ensemble techniques comes with the demand of reduction of noise in the predictive model. The ensemble technique uses multiple models and takes the average of them. This procedure helps to reduce the variances while controlling the bias. One of the preferred ensemble techniques is “Random Forest”.

Random forest was first proposed by Leo Breiman from the university of California in 2001.The random forest consists of multiple decision tress which are formed in such a way that each tree has no correlation with every other trees. These independent trees are used either for regression or classification problem.

The following are the steps used in Random Forest:

1. Fix a proper value of predictors usually labelled as “m”.
2. Select a new subset of predictors from the whole set of predictors depending on “m”
3. Train the data set with the feature subset to create decision tree for each group of training set
4. Choose a new and repeat the process above until travel all the feature subsets.
5. Test the prediction on the test data set.

Random forest consists of large number of decision trees. The randomness is incorporated in the model building process, including the selection of samples subset and feature subset which guarantees the independence of each decision tree. With the method of bagging where random boot strap samples are selected ensures the randomness of samples selection. Similarly, with the help of bagging subsets of features are selected based on the value of “m” from the entire feature set. The influence of “m’ is highly noted on the performance of random forest. With the increase in the value of “m”, the correlation between each tree in the training model can be improved to make the trees independent.

For our project, we fitted the random forest on each of the clusters and extracted the “variable importance” from four different models. Since, we had 11 predictors in all of our clustered data we set the range of “m” predictors to be for each cluster. The presence of small number of observations led us to use Leave One Out Cross-Validation folds for each cluster for deciding the optimal number of “m” predictors.

The number of optimal “m” was selected based on the corresponding RMSE. The table below shows the optimal number of “m” and corresponding RMSE on each cluster when LOOCV was implemented:

|  |  |  |
| --- | --- | --- |
| Cluster Number | Optimal “ m” tries | RMSE |
| 1 | 10 | 1.017949 |
| 2 | 6 | 2.400911 |
| 3 | 10 | 2.003527 |
| 4 | 8 | 2.451414 |

In R, “caret” package , for method = “rf” in “train(response~, data,method= “rf”, tune.grid=train\_control,tuneGrid=tune.grid)” was used to build the decision tree models on each of cross validation folds that resulted on an average model.

## Linear Regression

We wanted to see if there is any linear association between the response variable and the predictors. The ensemble technique above does not yield on a final single model. Therefore, we decided that it would be necessary to obtain a final model to make prediction on life expectancy based on the regressors we select. For this purpose, we considered multiple linear regression model.

Th model assumption for multiple linear regression are:

1. Errors are normally distributed
2. Constant variances
3. The relationship between predictors and response should be linear
4. The observations should be independent of each other

We often encounter the problem of model misspecification with the multiple linear regression. The model is prone to using extraneous variables which will create the problem of overfitting. It is also the case that the model lacks some important variables. In order to avoid these situations and obtain correctly specified models, we look for variable selection techniques. One technique that we used for variable selection is regularized regression technique. In a regularized regression technique, the regression coefficients are estimated by minimizing SSE+P, where SSE represents the sum of square deviations and P is the penalty factor. In regularized regression the estimated effects are slowly shrunken towards zero. The regularized regression has two types:

1. Ridge regression : minimizes SSE+
2. Lasso regression. Minimizes SSE+

is called the tuning parameter, and as the tuning parameter becomes very large, the penalty factor becomes large that forces the regression coefficients to become zero. This process results in reducing the variance, but bias is compromised unlike in ordinary least square regression.

In R, “glmnet” package is used for regularized regression. Since, regularized regression, requires the features to be standardized, “glmnet” package does the standardization for us. The parameter in the package was specified to use lasso regression by setting alpha=1.The is chosen based on the RMSE that corresponds to the minimum.

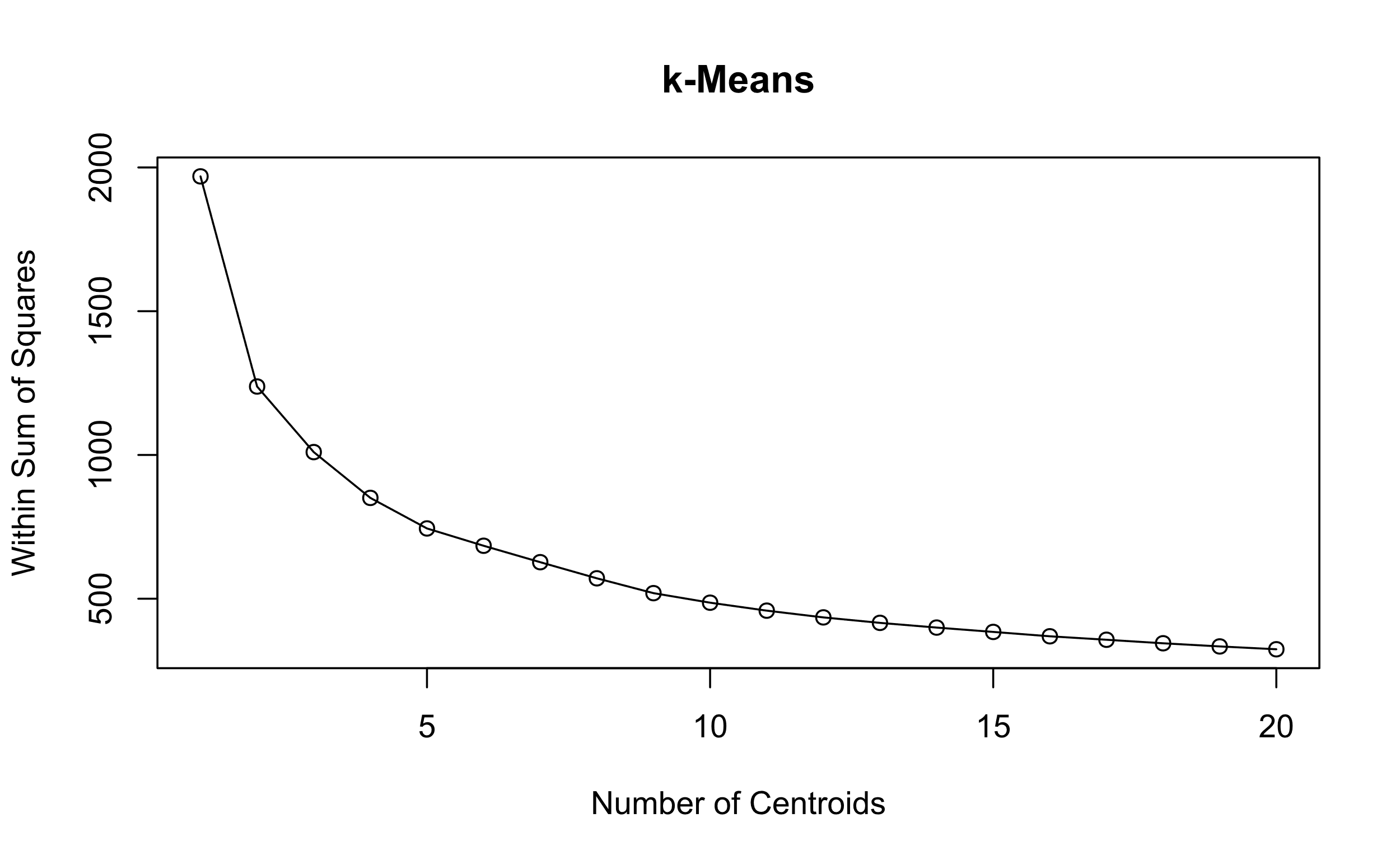
Once, the variables are selected the model is fitted with the appropriate regressors. It is also important that our model does not suffer from potential multicollinearity. So, once the model is fitted, we check the multicollinearity in the model using VIF (variance inflation factor) from the “car” package. The VIF values higher than 4 signifies that the model suffers from moderate multicollinearity, whereas the VIF higher than 8 signifies that the model suffers from severe multicollinearity. If the problem of multicollinearity exists, then we will drop the highly correlated variables to get rid of the multicollinearity from the existing model. The model without multicollinearity is used for the prediction purposes. The significance of the regressors in the model will be compared with the significance level of =0.05.

# Results

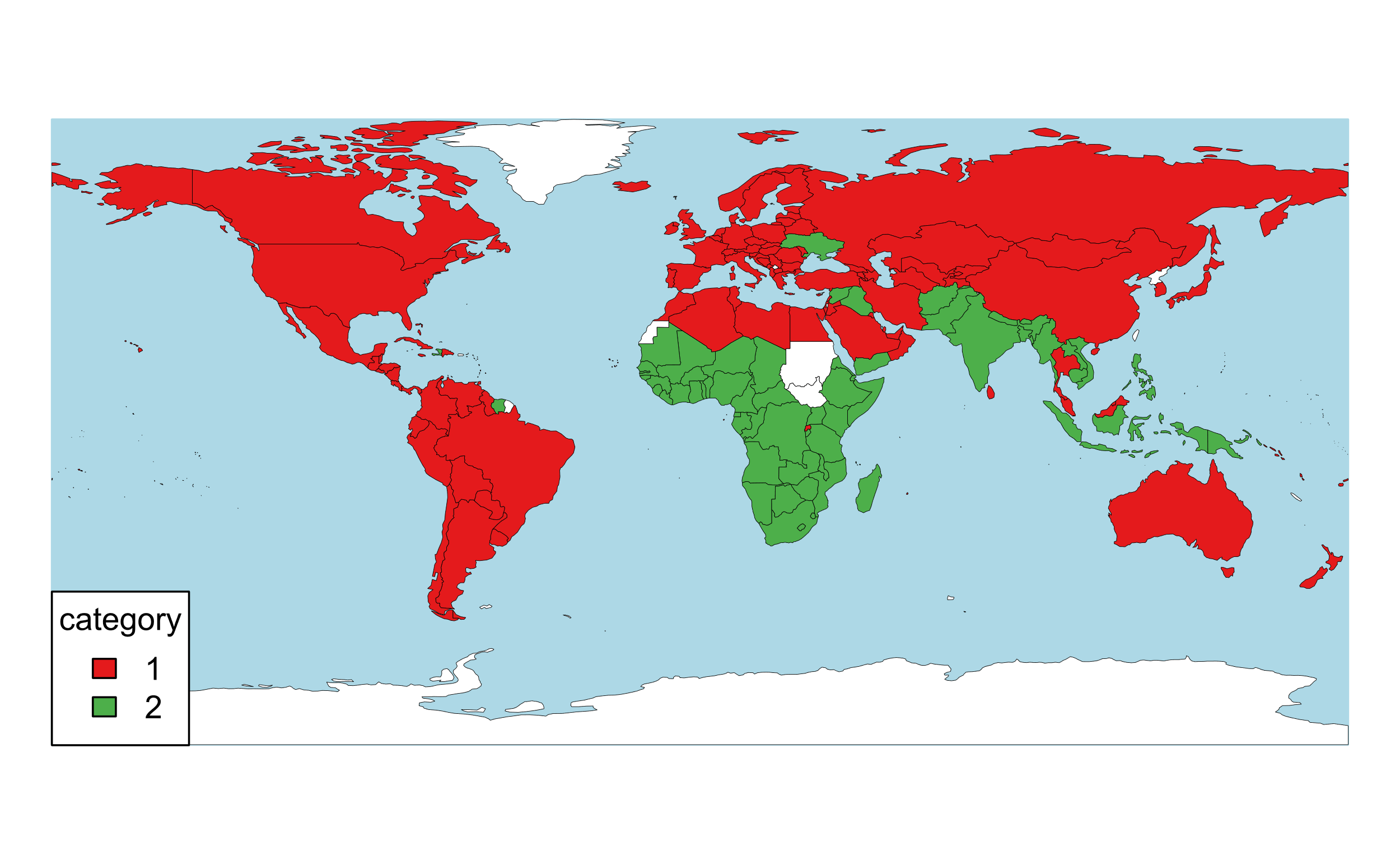
## k-NN

## k-Means

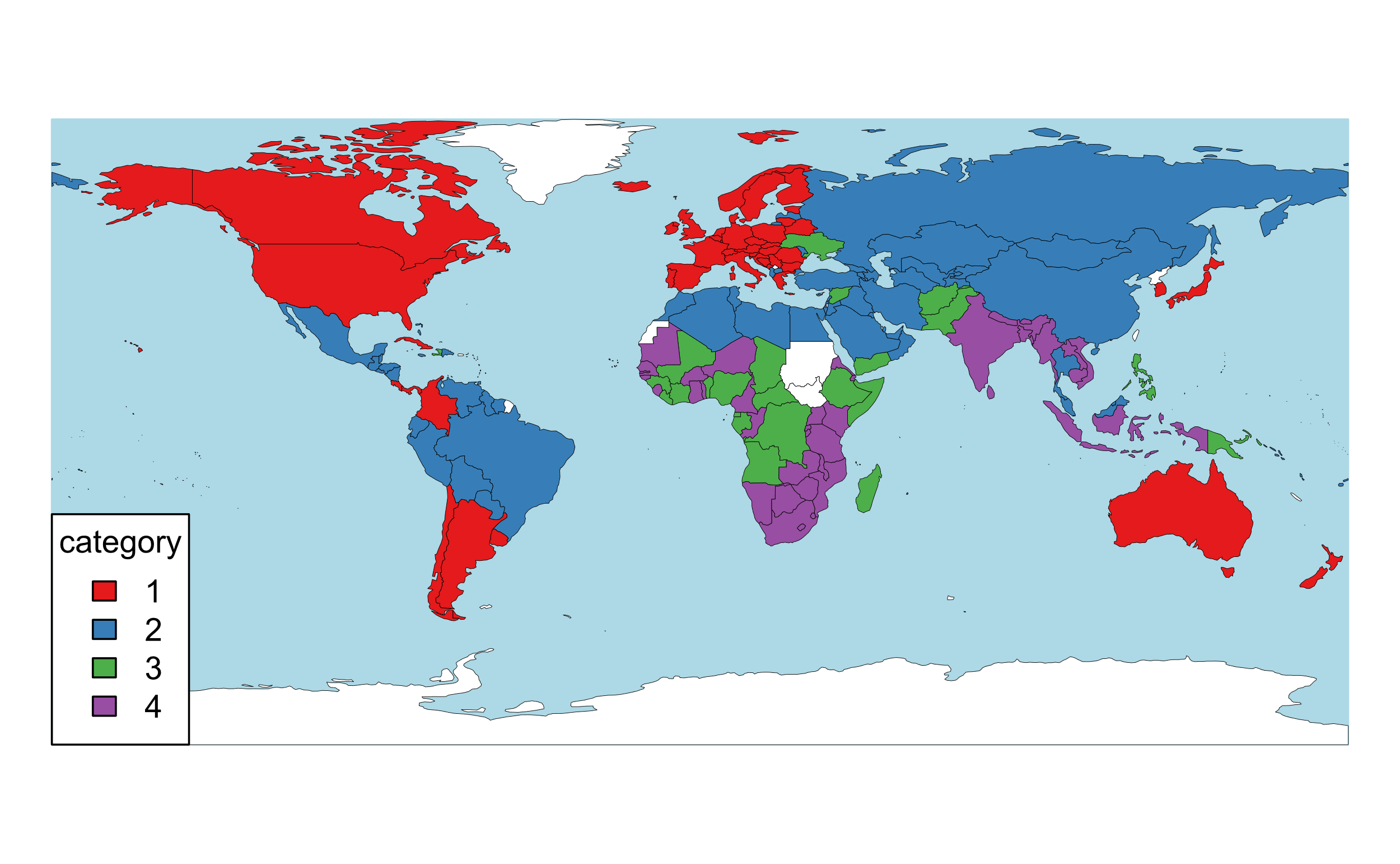
Following our methods in section 3.3, we ran the k-means algorithm using various values of using the same set of 1000 seeds for each . The results are displayed in the figure below.



We can see that there is not a distinct “elbow” in the graph that can be chosen as our optimal other than a slight bend at = 2 which results in the map seen below.



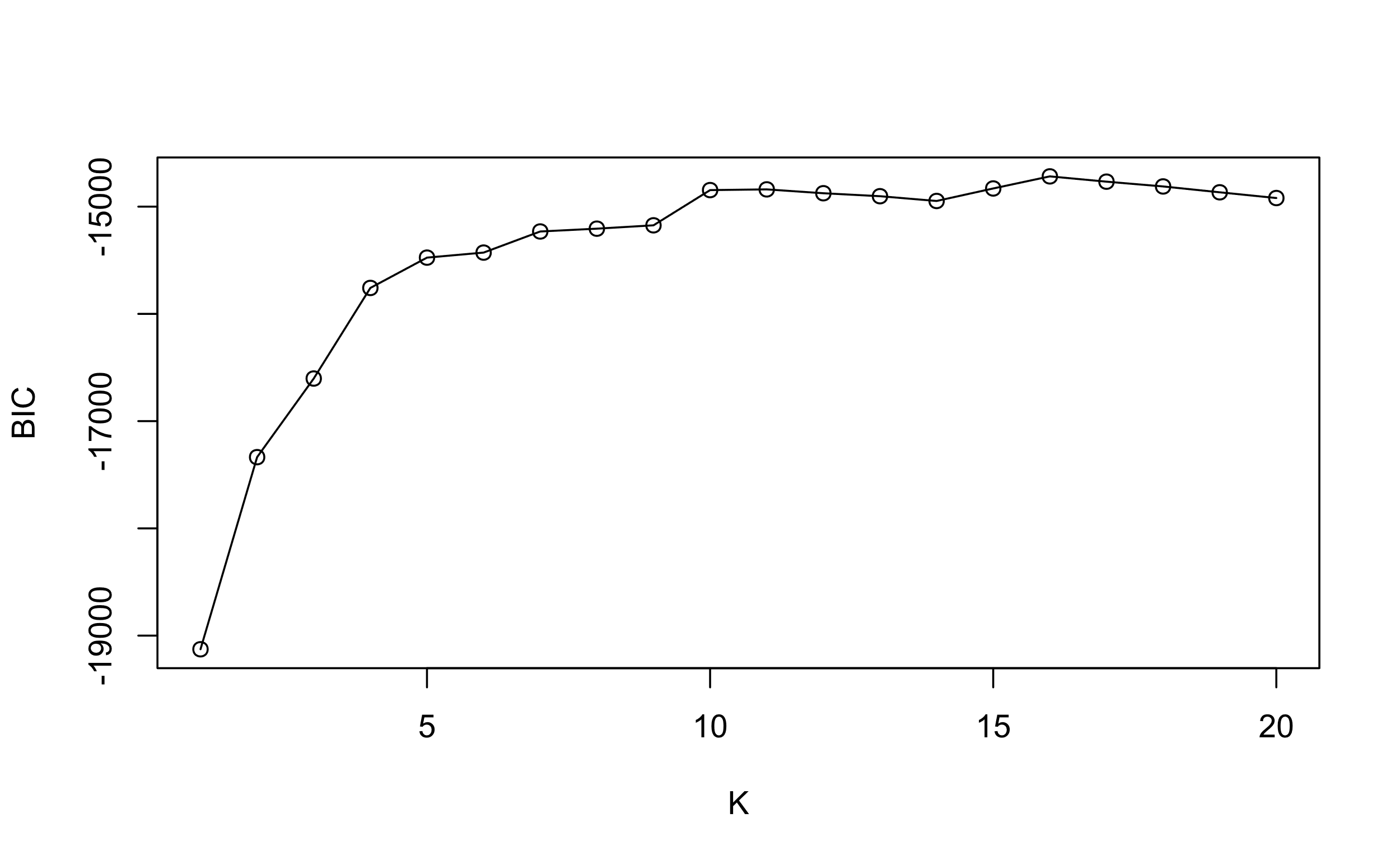
The algorithm did a decent job at clustering the countries; however, there are issues that come with grouping the countries in such a broad manner. For example, less economically developed countries such as Turkey and Ecuador are grouped with more economically developed countries such as Germany and South Korea. We looked at = 4 to help break up some of these oddities. We chose = 4 because reduction in the within sum of squares beyond this point was negligible. The resulting map can be seen below.



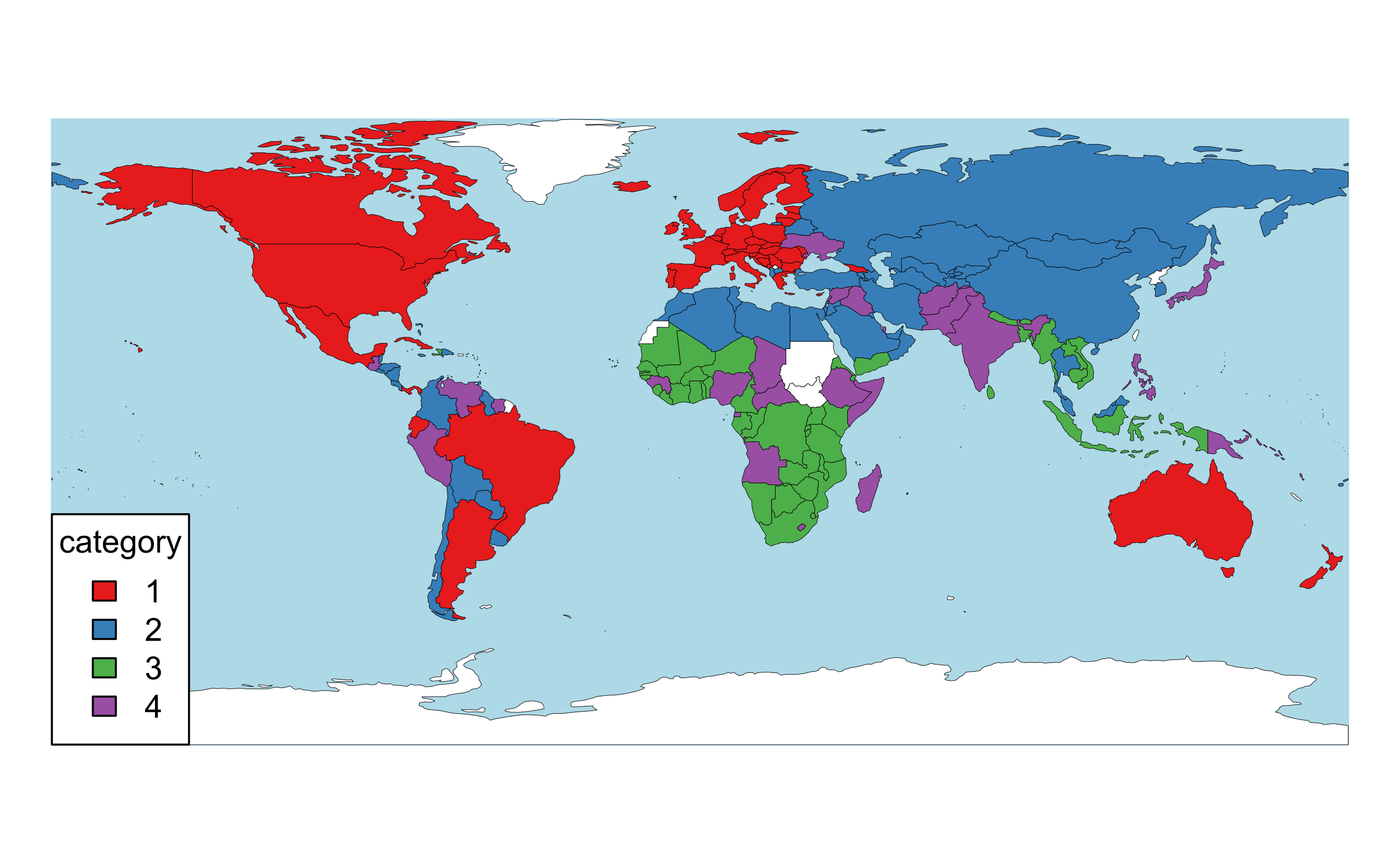
We can see that the k-means algorithm has done a good job of splitting up the African continent in a way that makes sense while keeping Japan and South Korea grouped with countries like the America and Australia.

## Model-Based

Using the methods described in section 3.4 we computed the computed the negative BIC of the model-based clustering algorithms as various values of . The results are shown below.



As with the k-means model there is not a clear, reasonable optimal number of components, so we will be choosing 4 components for this model as the gain in negative BIC from adding more components is negligible. The resulting clusters can be seen in the map below.



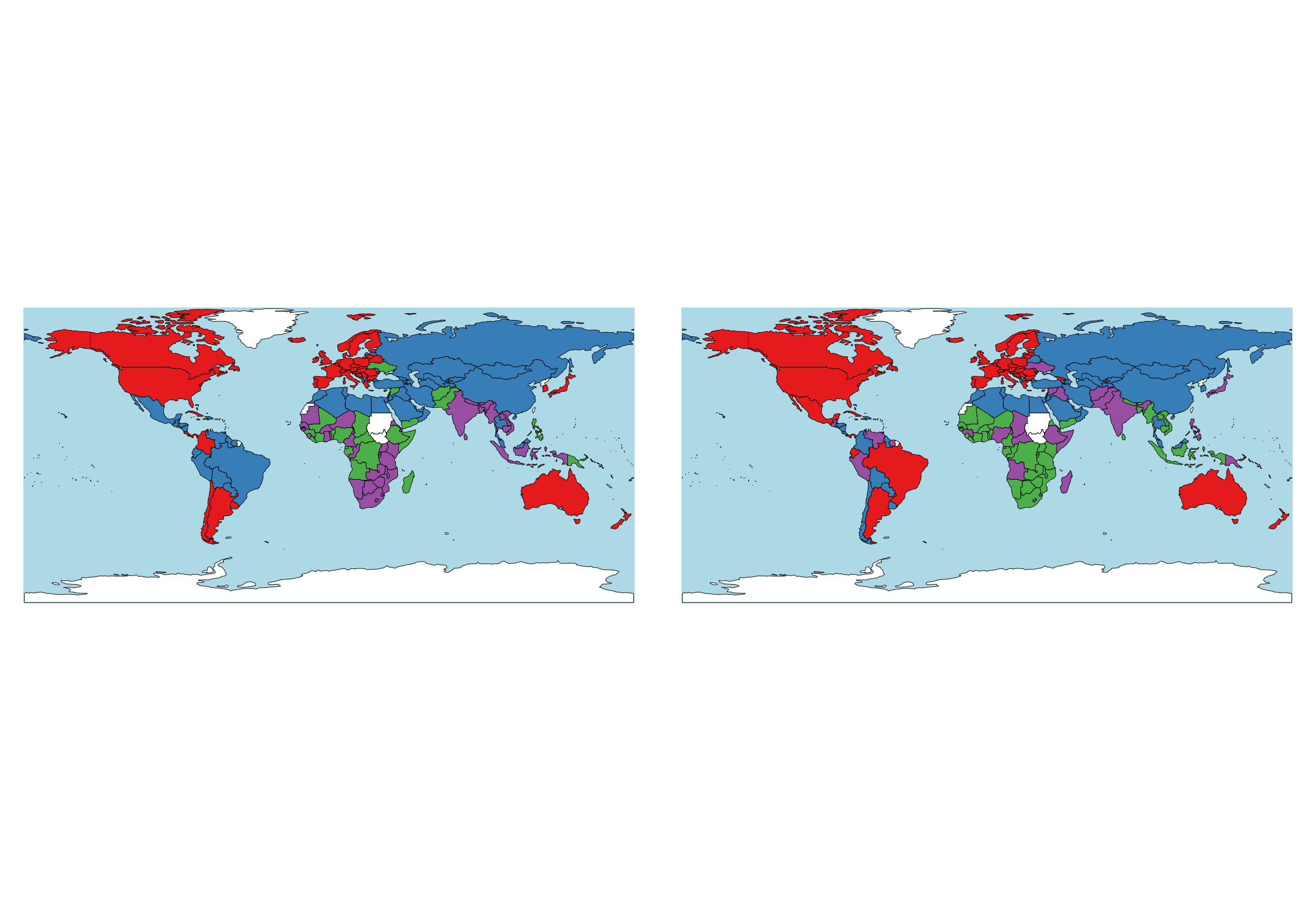
We can see that the model-based clustering did well but there are some problems, specifically in Asia where Japan is grouped with India and South Korea is grouped with Russia.

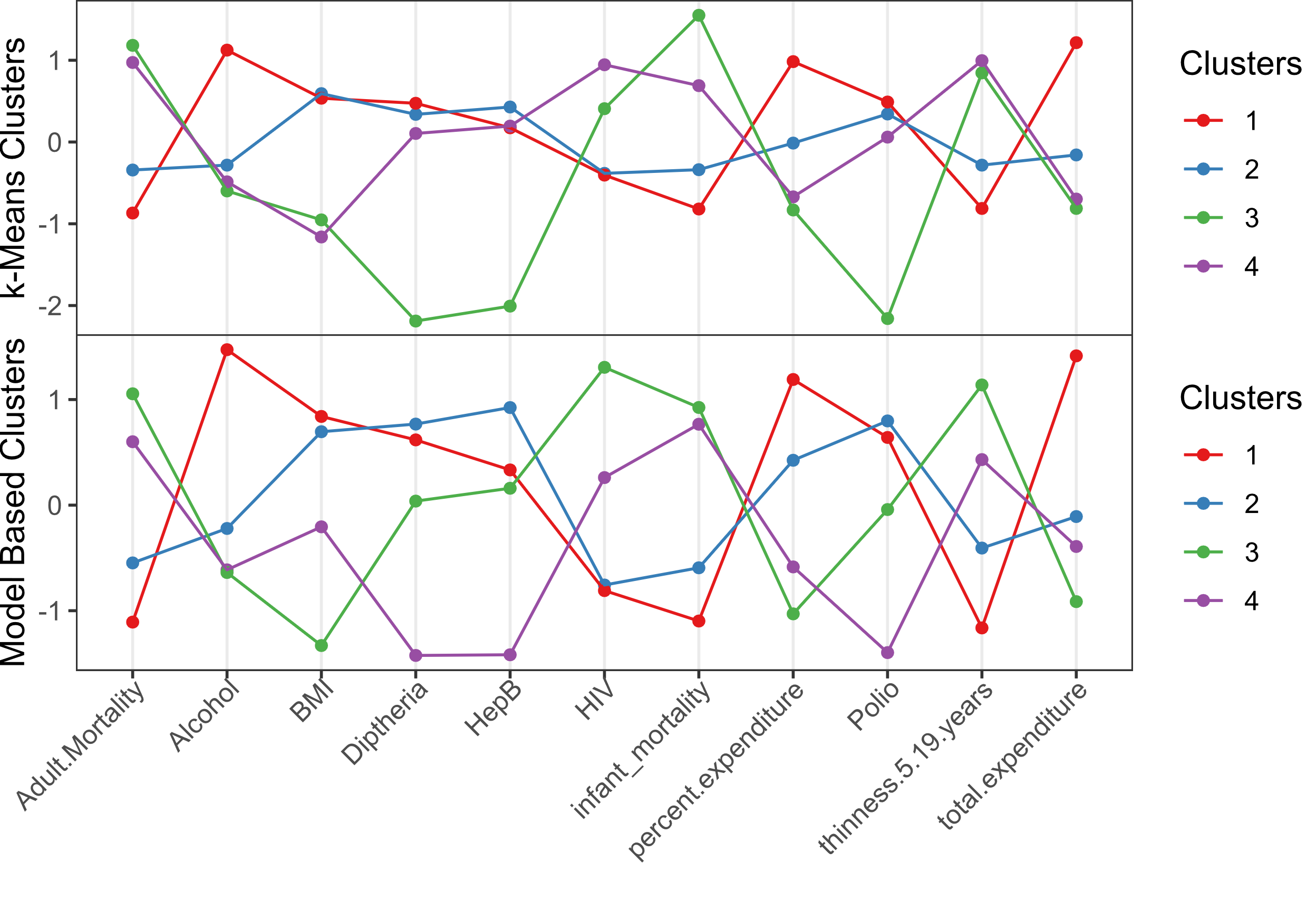
## K-Means vs Model-Based

We believe that, visually, the 4-means clustering method did a better job splitting up Africa and Eastern Asia, specifically with South Korea and Japan in Asia, and South Africa and Somalia in Africa, as seen in the maps below. This is backed up by the scaled means below. For example, Japan should not be in the group with countries that have the lowest Hepatitis B and Diphtheria vaccination rates. For this reason, we decided to choose the k-means clustering model.

Model Based Clusters

k-Means Clusters

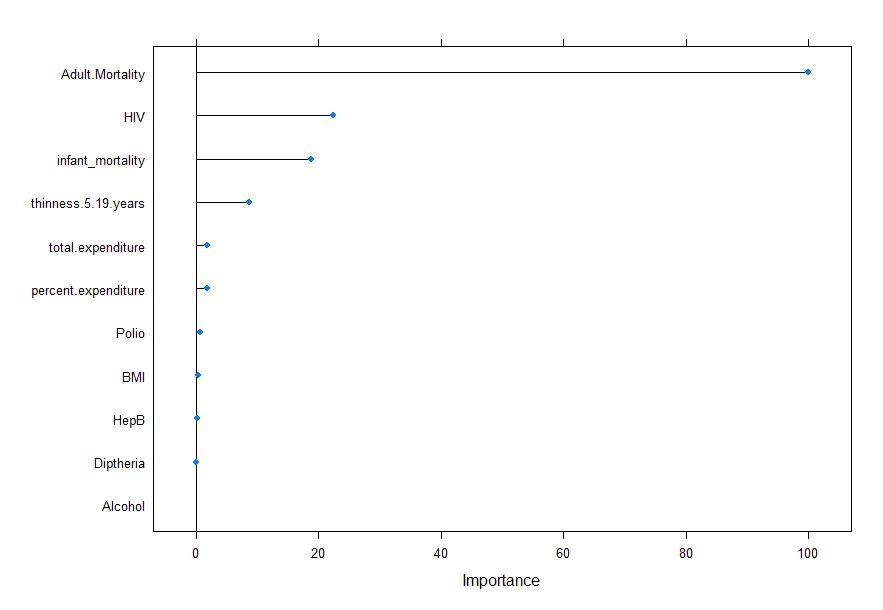




## Random Forest

### Cluster 1

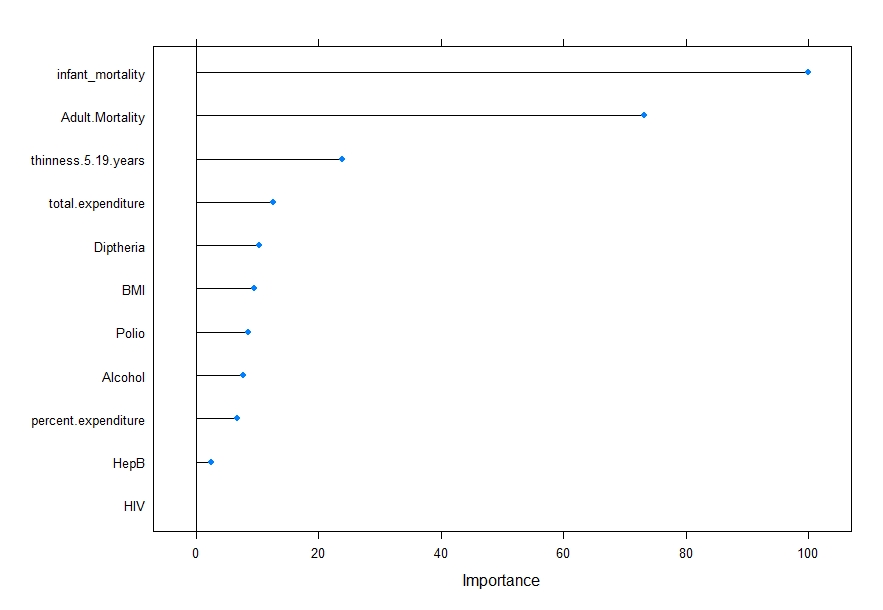
After random forest was fitted into cluster 1, it was found that “Adult mortality” was ranked as a powerful predictor outranking all other variables followed by “HIV”, and “Infant Mortality” etc. The complete ranking of the variable importance is given below:



The variables ranked in Cluster 1 informs us that the variable Adult mortality is very important in explaining the life expectancies at birth of the populations in the cluster. HIV and Infant mortality are also important features to be considered. Based on the data from the year 2014, the countries in the cluster 1 would need to extensively work on improving the physical, mental well-being of adult population. Monitoring the risk of HIV among the youths would be important and malnutrition among the children should be taken seriously to improving the life expectancy

### Cluster 2

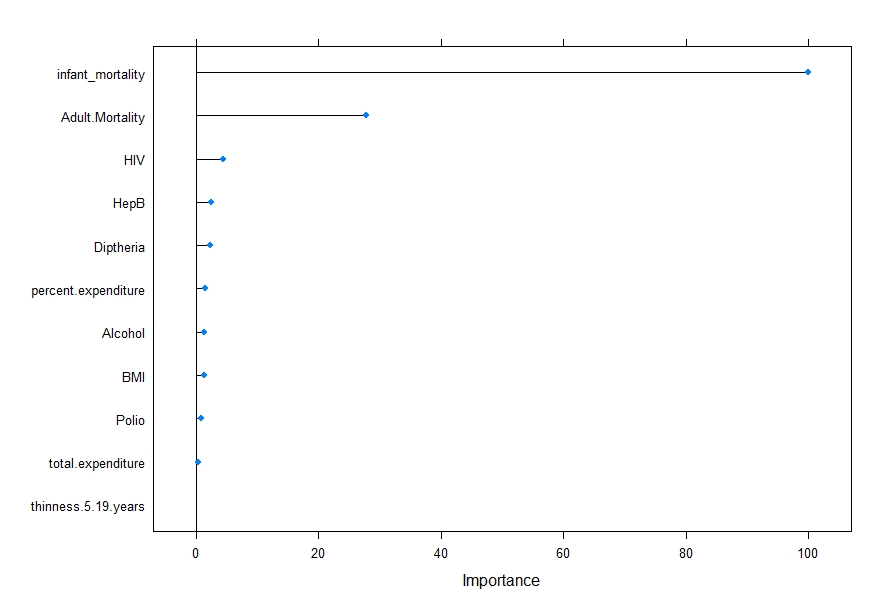
The similar approach was used for cluster 2. Upon inspection of the variable importance, it was found that “infant mortality” was ranked as a powerful predictor outranking all other variables closely followed by “adult mortality”. The complete ranking of the variable importance is given below:



The variables ranked tells us that the countries in Cluster 2 need to focus on high mortality rate among the population. The thinness among children for age group between 5 to 19 years is a factor that needs to taken into consideration. The government and other institutions should increase percent of the expenditure spent on the population to alleviate the problems.

4.5.3 Cluster 3

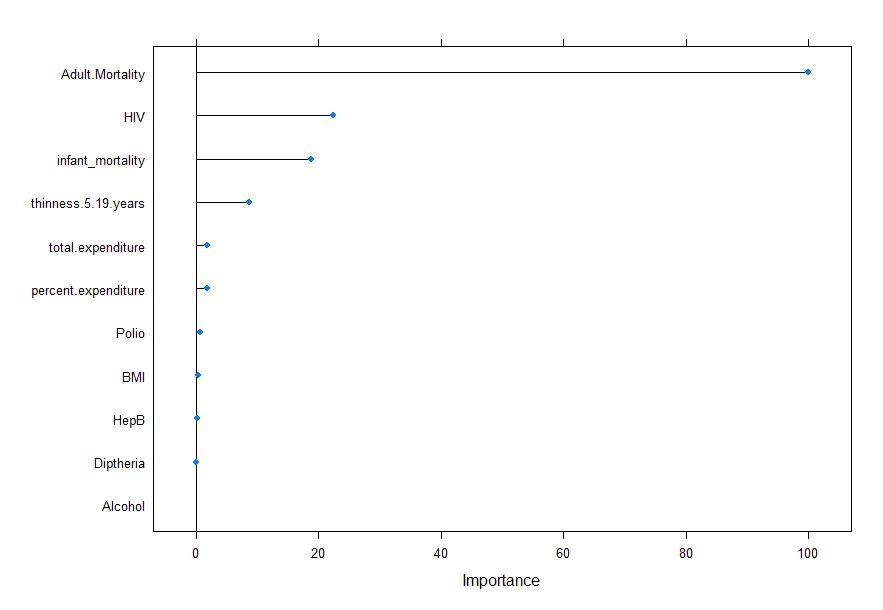
For cluster 3 , “infant mortality” was ranked as a powerful predictor outranking all other variables closely followed by “adult mortality”. “HIV” . The complete ranking of the variable importance is given below



. The ranking of first two variables is similar in cluster 2. However, thinness among children aged 5 to 19 years seemed to have no contribution in the prediction unlike in cluster 2. Countries in cluster 3 should also heavily focus on the two features “Adult Mortality” and “Infant Mortality”. It is also recommended to consider the features such as HIV, Hepatitis B and Diphtheria to make significant improvement in the life expectancy.

4.5.4 Cluster 4

For cluster 4, “Adult mortality” was ranked as a powerful predictor outranking all other variables followed by “HIV”, “infant mortality, and “thinness.5.19. years” etc. The complete ranking of the variable importance is given below



Countries in cluster 4 share the same first three predictors as in cluster 1 and 3. Countries in this cluster should follow the similar approach as mentioned earlier for cluster 1 and cluster 3.

The comparison table of variable importance among different cluster is given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | Cluster1 | Cluster 2 | Cluster 3 | Cluster 4 |
| 1 | Adult Mortality | Infant Mortality | Infant Mortality | Adult Mortality |
| 2 | HIV | Adult Mortality | Adult Mortality | HIV |
| 3 | Infant Mortality | Thinness (5-19) | HIV | Infant Mortality |
| 4 | Thinness (5-19) | Total Expenditure | Hepatitis B | Thinness (5-19) |
| 5 | Total Expenditure | Diphtheria | Diphtheria | Total Expenditure |
| 6 | Percent Expenditure | BMI | Percent Expenditure | Percent Expenditure |
| 7 | Polio | Polio | Alcohol | Polio |
| 8 | BMI | Alcohol | BMI | BMI |
| 9 | Hepatitis B | Percent Expenditure | Polio | Hepatitis B |
| 10 | Diphtheria | Hepatitis B | Total Expenditure | Diphtheria |
| 11 | Alcohol | HIV | Thinness (5-19) | Alcohol |

4.6 Linear Regression

The method described in section 3.6 was used for each cluster.

4.6.1 Cluster 1

For the cluster 1, reasonable transformations were made on the predictors “Alcohol”, “HepB”, “infant\_mortality” and percent.expenditure”, as these variables had high deviations from the linearity assumptions. Also, we needed to achieve a linear relationship between the regressors and the response variable “HALE\_Birth”. The variance inflation factor of the predictors indicated the model was suffering from multicollinearity. Lasso regression was carried out on the model and “Diphtheria” was shrunk to zero. The final model had an Adjusted R2 value of 0.9042 and residual standard error value of 0.6593. The significant predictors in the final model were “Adult.Mortality”, “BMI”, “Polio”, “logpercent\_expenditure”, “thinness.5.19.years” and “total.expenditure” as shown in the final model in appendix A.

4.6.2 Cluster 2

A similar approach was carried out for cluster 2. Transformations were made on “Alcohol”, “HepB”, “infant\_mortality”, percent.expenditure”, “thinness.5.19.years” and “total.expenditure”. The model indicated the presence of multicollinearity. Lasso regression shrunk “Diphtheria”, “BMI” and “HepB\_power\_5” to zero. The final model had an Adjusted R2 value of 0.7102 and a residual standard error value of 1.939. The significant predictors in the final model were “Adult.Mortality”, “logAlcohol”, “logpercent\_expenditure” and “logInfant” as shown in the final model in appendix B.

4.6.3 Cluster 3

Transformations were made on “Alcohol”, “HepB”, “infant\_mortality”, percent.expenditure” and “total.expenditure”. The model indicated the presence of multicollinearity. Lasso regression shrunk “BMI”, “Polio”, “logpercent\_expenditure” and “thinness.5.19.years” to zero. The final model had an Adjusted R2 value of 0.8329 and a residual standard error value of 1.7. The only significant predictor in the final model was “logInfant” as shown in the final model in appendix C.

4.6.4 Cluster 4

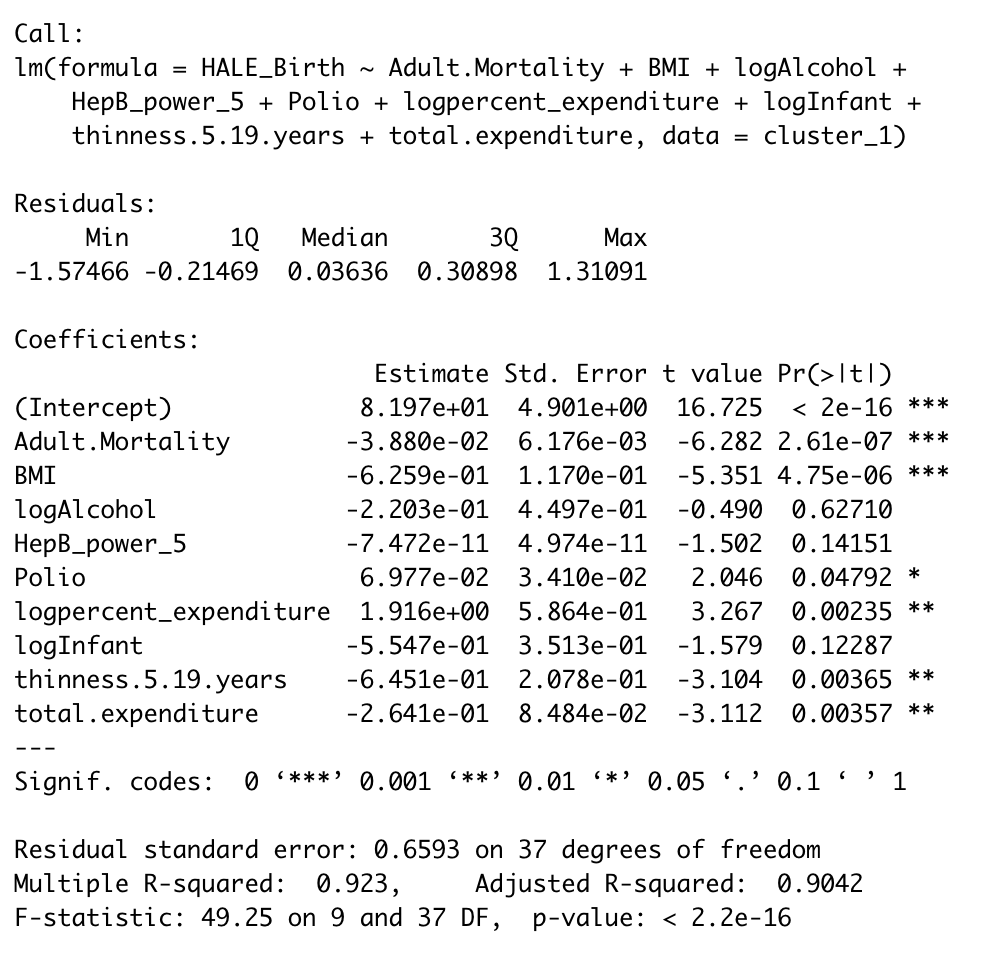
Transformations were made on “Alcohol”, “HepB” and “total.expenditure”. The model indicated the presence of multicollinearity. Lasso regression shrunk “Diphtheria”, “logAlcohol”, “Polio”, “thinness.5.19.years” and “logtotal.expenditure” to zero. The final model had an Adjusted R2 value of 0.9029 and a residual standard error value of 1.564. The significant predictors in the final model were “Adult.Mortality” and “HIV” as shown in the final model in appendix D.

# Conclusion

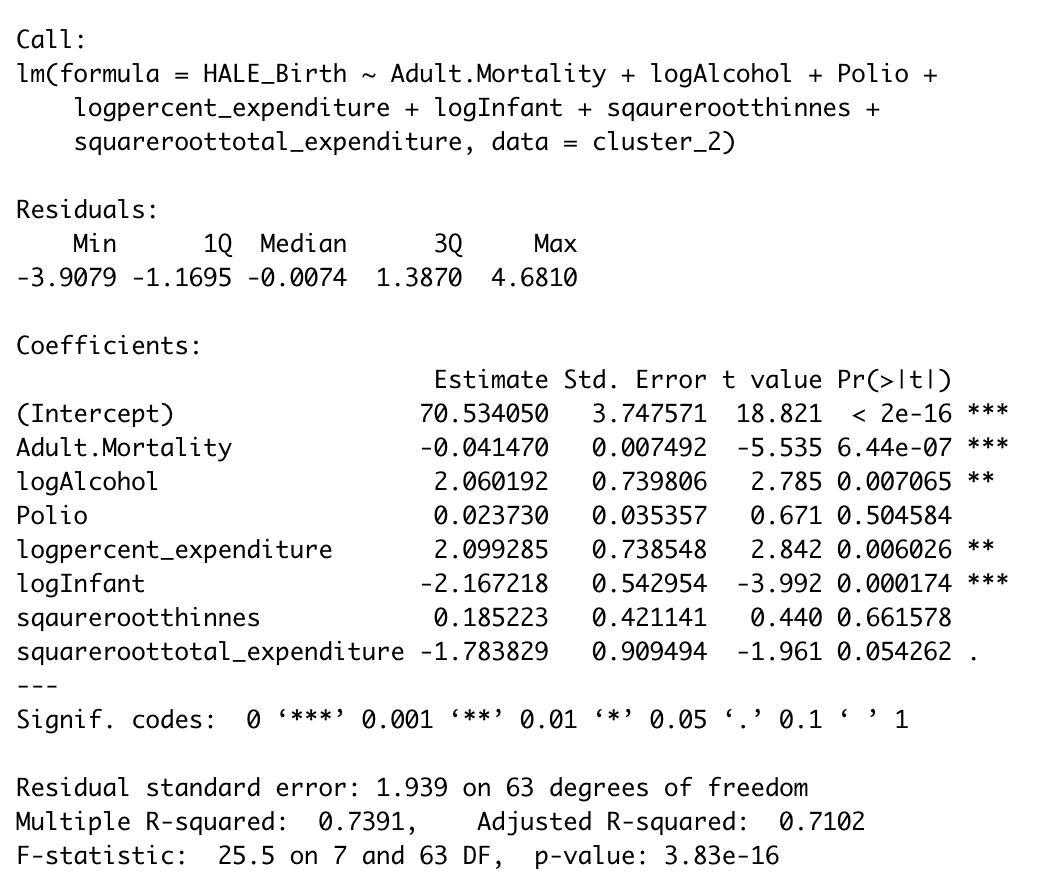
“The report should conclude with a conclusion section where main takeaways, limitation of the analysis and future direction should be discussed briefly.”

# 6 Appendix

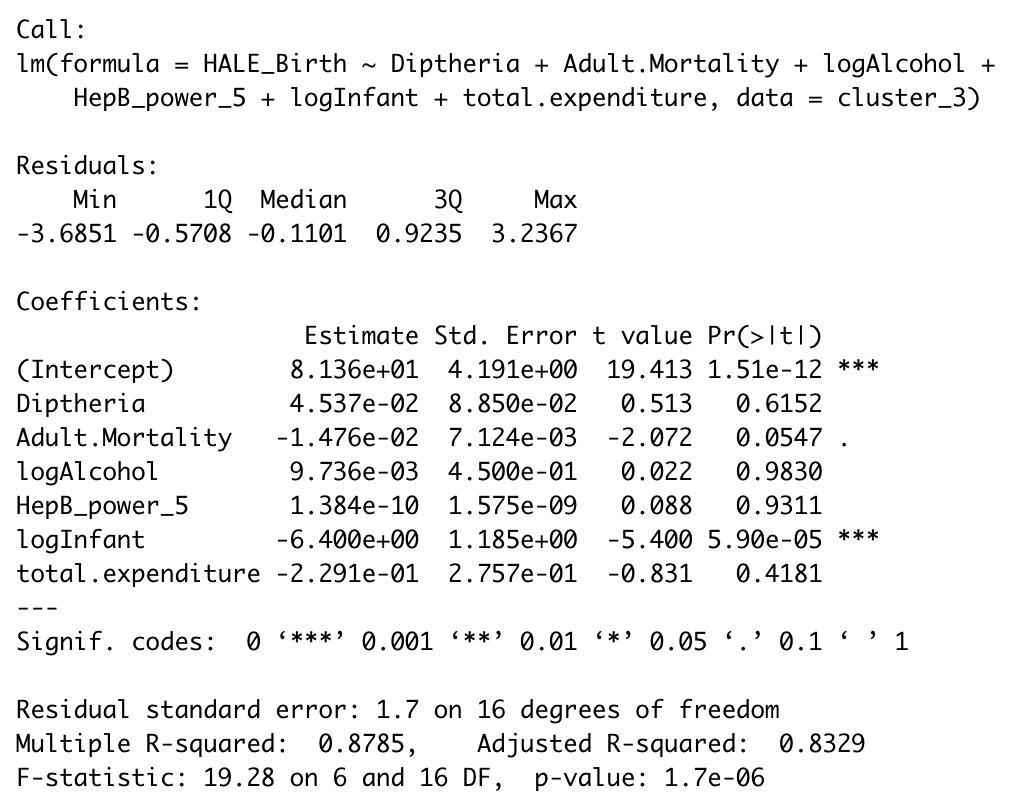
**Appendix A: Final Model for Cluster 1**



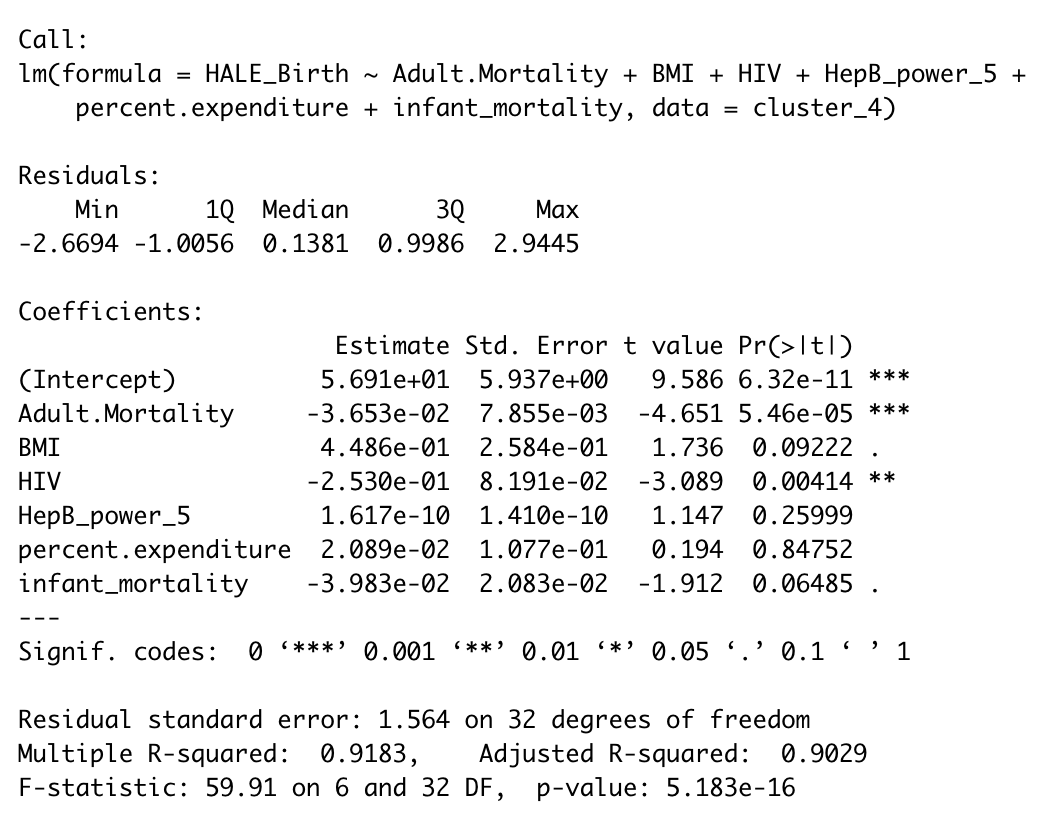
**Appendix B: Final Model for Cluster 2**



**Appendix C: Final Model for Cluster 3**



**Appendix D: Final Model for Cluster 4**



**Table: Significant Predictors for the Clusters**

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| Adult.Mortality | Adult.Mortality | logInfant | Adult.Mortality |
| BMI | logAlcohol |  | HIV |
| Polio | logpercent\_expenditure |  |  |
| logpercent\_expenditure | logInfant |  |  |
| thinness.5.19.years |  |  |  |
| total.expenditure |  |  |  |

# 7 Citations

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., & Li, S. (2019). FNN: Fast Nearest Neighbor Search Algorithms and Applications. *R package version 1.1.3*. Retrieved from https://CRAN.R-project.org/package=FNN

Kuhn, M. (2021). caret: Classification and Regression Training. R package version 6.0-90. Retrieved from https://CRAN.R-project.org/package=caret

Parmar, A., Katariya, R., & Patel, V. (2018). A review on random forest: An ensemble classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things* (pp. 758-763). Springer. doi:10.1007/978-3-030-03146-6\_86

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

RStudio Team. (2021). RStudio: Integrated Development Environment for R. Boston, MA. Retrieved from http://www.rstudio.com/

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*(1), 289-317. doi:https://doi.org/10.32614/RJ-2016-021

Torgo, L. (2010). Data Mining with R, learning with case studies. Chapman and Hall/CRC. Retrieved from http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR